Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum.

Abstract

Importance

The rapid expansion of virtual health care has caused a surge in patient messages concomitant with more work and burnout among health care professionals. Artificial intelligence (AI) assistants could potentially aid in creating answers to patient questions by drafting responses that could be reviewed by clinicians.

Objective

To evaluate the ability of an AI chatbot assistant (ChatGPT), released in November 2022, to provide quality and empathetic responses to patient questions.

Design, Setting, and Participants

In this cross-sectional study, a public and nonidentifiable database of questions from a public social media forum (Reddit's r/AskDocs) was used to randomly draw 195 exchanges from October 2022 where a verified physician responded to a public question. Chatbot responses were generated by entering the original question into a fresh session (without prior questions having been asked in the session) on December 22 and 23, 2022. The original question along with anonymized and randomly ordered physician and chatbot responses were evaluated in triplicate by a team of licensed health care professionals. Evaluators chose "which response was better" and judged both "the quality of information provided" (very poor, poor, acceptable, good, or very good) and "the empathy or bedside manner provided" (not empathetic, slightly empathetic, moderately empathetic, empathetic, and very empathetic). Mean outcomes were ordered on a 1 to 5 scale and compared between chatbot and physicians.

Results

Of the 195 questions and responses, evaluators preferred chatbot responses to physician responses in 78.6% (95% CI, 75.0%-81.8%) of the 585 evaluations. Mean (IQR) physician responses were significantly shorter than chatbot responses (52 [17-62] words vs 211 [168-245] words; t = 25.4; P < .001). Chatbot responses were rated of significantly higher quality than physician responses (t = 13.3; P < .001). The proportion of responses rated as good or very good quality (≥ 4), for instance, was higher for chatbot than physicians (chatbot: 78.5%, 95% CI, 72.3%-84.1%; physicians: 22.1%, 95% CI, 16.4%-28.2%;). This amounted to 3.6 times higher prevalence of good or very good quality responses for the chatbot. Chatbot responses were also rated significantly more empathetic than physician responses (t = 18.9; P < .001). The proportion of responses rated empathetic or very empathetic (≥ 4) was higher for chatbot than for physicians (physicians: 4.6%, 95% CI, 2.1%-7.7%; chatbot: 45.1%, 95% CI, 38.5%-51.8%; physicians: 4.6%, 95% CI, 2.1%-7.7%). This amounted to 9.8 times higher prevalence of empathetic or very empathetic responses for the chatbot.

Conclusions

In this cross-sectional study, a chatbot generated quality and empathetic responses to patient questions posed in an online forum. Further exploration of this technology is warranted in clinical settings, such as using chatbot to draft responses that physicians could then edit. Randomized trials could assess further if using AI assistants might improve responses, lower clinician burnout, and improve patient outcomes.